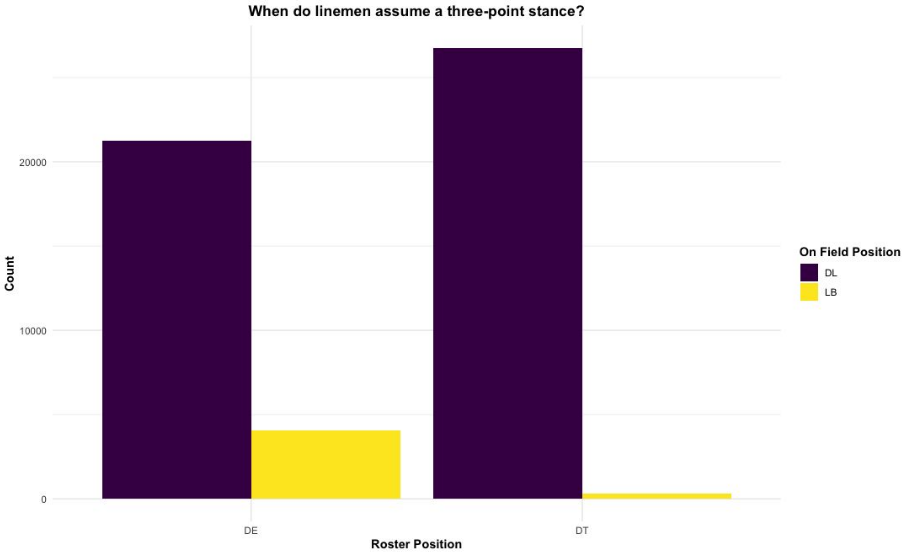
SIS Football Analytics Challenge

Alex Stern (University of Virginia)

Positional Consideration



- ➔ Those listed on the roster as linemen, usually line up as linemen
- ➔ Using **roster position** was the *most* clear-cut way of dividing players
- ➔ NFL Draft, Free Agency, etc. almost exclusively refer to roster position (unless you're Isaiah Simmons)

Other Assumptions

- ➔ All spike plays were filtered out immediately

Feature Engineering

- isHome: teams were matched up with their respective home stadiums
- gapMatch: on running plays, is the lineman matched up with the intended gap (ex. If the RB goes to the Right B Gap; the lineman’s technique is 2, 3, 4i, or 4; and he is on the left side of the ball (defensive perspective), then he is matched up with the RB’s intended gap

Feature Engineering cont.

- ➔ nDT: # of defensive tackles on the play
- ➔ nDE: # of defensive ends on the play
- ➔ pass_epa: EPA per pass attempt defended over the entire data set by the defensive team
- ➔ rush_epa: EPA per rush attempt defended over the entire data set by the defensive team

*The last two help control for possible strengths and resulting relationships in play calling decisions (Yurko, Ventura, Horowitz, 2018)

Multilevel Modeling (Random Effects)

- ➔ “Since every play involving the same player is a repeated measure of performance, the plays themselves are not independent.” (Yurko, Ventura, Horowitz, 2018)
- ➔ The positive/negative and “roughly symmetric” distribution of EPA lends itself to a helpful normality assumption
- ➔ Each lineman’s avg. effect can be interpreted as their individual points added (iPA)

Multilevel Modeling cont.

- ➔ Linemen “involved in fewer plays will be pulled toward the group mean level (0 in this case) as compared to those involved in more plays and thus carrying more information, resulting in partially pooled estimates.” (Gelman and Hill, 2007)
- ➔ Models are fit using penalized likelihood (Bates et al., 2015)

Multilevel Modeling cont.

- ➔ Turning pressures into sacks/turnovers is not a repeatable skill. (Riske, 2020)
- ➔ Pressures are a far more stable stat YOY
- ➔ Mean EPA on passing plays with pressure (-0.4) vs. without pressure (0.2) is significantly different
- ➔ Turnovers (fumbles, interceptions) account for large EPA swings -> will be accounted for as covariates
- ➔ Thus, the model will attempt to decipher who is delivering iPA in the form of *pressures*

Passing Model

$$EPA_i \sim \mathcal{N}(D_{d[i]} + F_{f[i]} + \mathbf{P_i} \cdot \mathbf{p}, \sigma_{EPA_i}) \text{ for } i = 1, ..., n \text{ plays,}$$

$$D_d \sim \mathcal{N}(\mu_D, \sigma_D^2) \text{ for } d = 1, ..., \text{ \# of defensive linemen,}$$

$$F_f \sim \mathcal{N}(\mu_F, \sigma_F^2) \text{ for } f = 1, ..., \text{ \# of offenses}$$

The covariate vector **P**$_{\mathbf{i}}$ contains a set of indicator variables for isHome, nDT, nDE, Turnover, FumbleByReceiver, ThrowDepth, Completion, and run_epa. **p** is the corresponding coefficient vector.

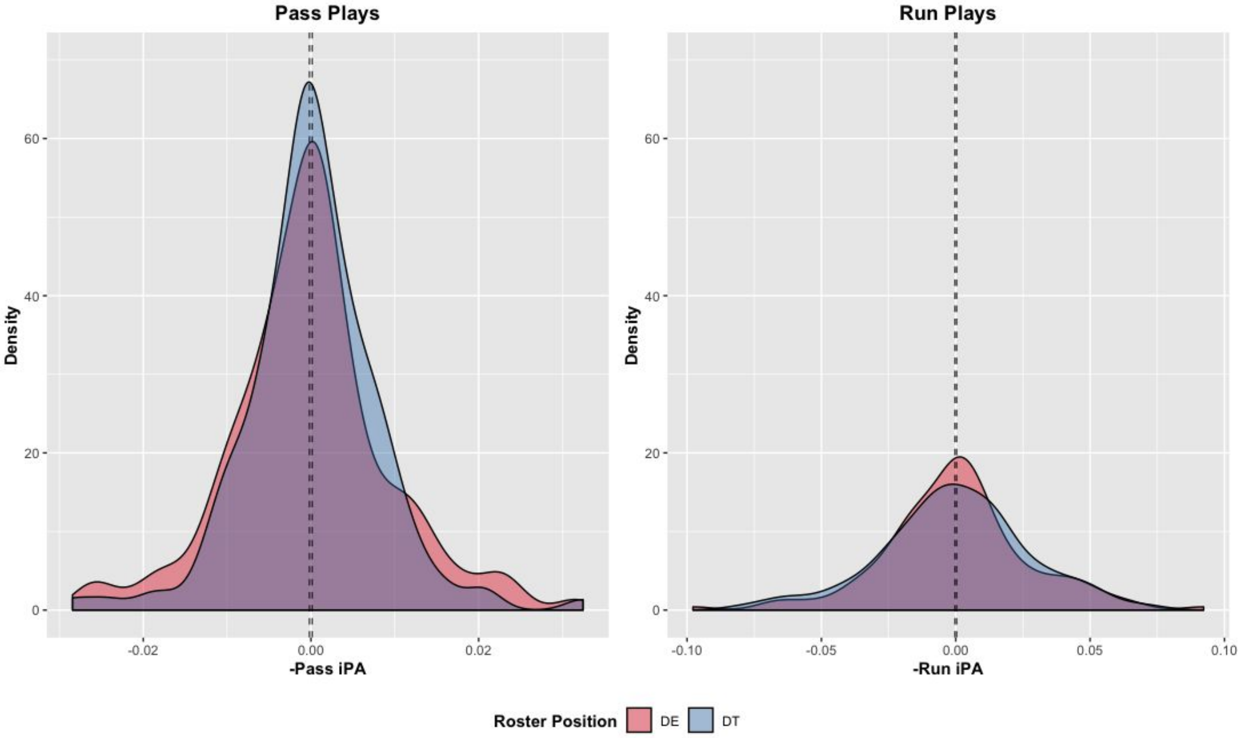
Rushing Model

$$EPA_i \sim \mathcal{N}(D_{d[i]} + T_{t[i]} + F_{f[i]} + \mathbf{R_i} \cdot \mathbf{r}, \sigma_{EPA_i}) \text{ for } i = 1, ..., n \text{ plays},$$
$$D_d \sim \mathcal{N}(\mu_D, \sigma_D^2) \text{ for } d = 1, ..., \text{ \# of defensive linemen},$$
$$T_t \sim \mathcal{N}(\mu_T, \sigma_T^2) \text{ for } t = 1, ..., \text{ \# of team-side-gaps},$$
$$F_f \sim \mathcal{N}(\mu_F, \sigma_F^2) \text{ for } f = 1, ..., \text{ \# of offenses}$$

The covariate vector **R**$_{\text{i}}$ contains a set of indicator variables for isHome, gapMatch, nDT, nDE, Turnover, and pass_epa. **r** is the corresponding coefficient vector.

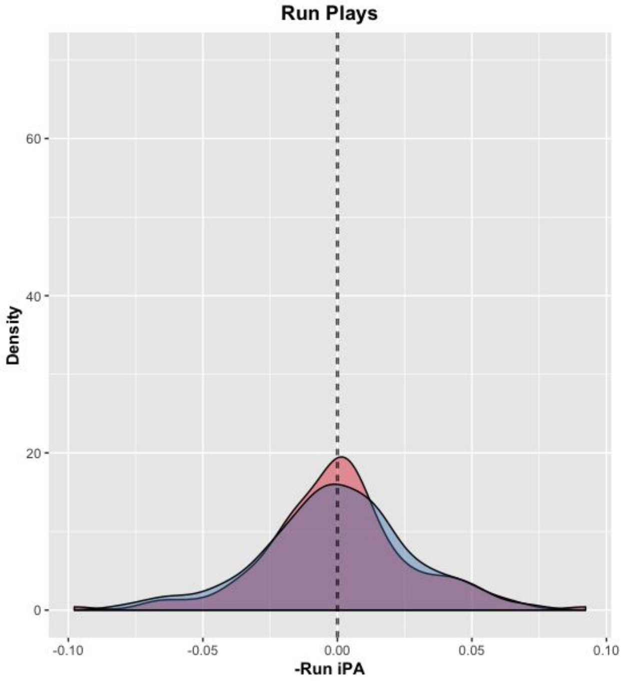
So, which is the most valuable defensive line position? Well…

...it depends



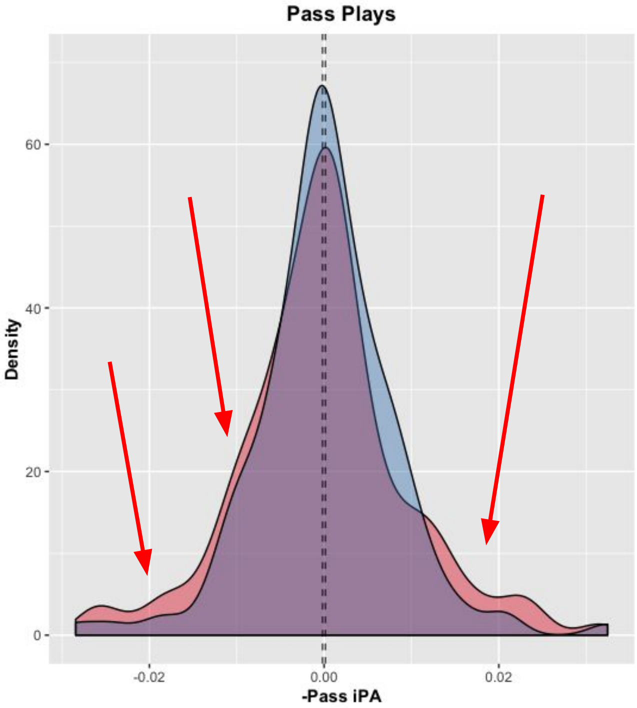
Distribution of (-1 * iPA) values, since the goal of the defense is to reduce the expected points of an offenses' drive

Analysis



- ➔ Run plays provide very little information about the differing abilities b/t DE and DT
- ➔ As many Big Data Bowl solutions pointed out, most information about the distribution of EPA on running plays is heavily dependent on the # of men in the box
- ➔ The gapMatch covariate also provided a significant amount of information based on model results

Analysis cont.



- ➔ It’s clear on passing plays that the distribution of talent among defensive ends has a **larger variance** (constant overlap, longer/thicker tails)
- ➔ This implies that securing an elite player at the **DE** position is **more valuable** than doing the same for the DT position
- ➔ The talent discrepancy b/t a 70th and 80th percentile DE is far greater than that of a DT

Where could this change?

- ➔ PFF’s Timo Riske, Eric Eager, and many others have pointed out the importance of QB pressure (Riske, 2020)
- ➔ On 3rd down, QB’s are more likely to hold on to the ball for longer in order to secure their read since throwing it away/scrambling is often not an option
- ➔ P(QB Pressure) is inherently a (likely exponential) function of time to throw
- ➔ Yards to go for a first down likely has an effect both on play calling and play anticipation by the defense

Analysis cont.

- ➔ Unfortunately, consecutive filters for specific down and distance situations led to sample sizes too small for a model this complex to run to fruition on (see Limitations and Future Analysis)
- ➔ In situations where the model was able to converge (all mid/long downs, 1st/2nd down only, etc.) there were no marked differences from the size, shape, and relative overlap of the density curves shown earlier -> DE still the more valuable roster position

Limitations and Future Analysis

- With a season-long dataset, minute trends based on game situation could possibly become more apparent and I'd be interested to know which (if any) are significant in relation to this problem/model
- With multiple years of data, the season-to-season stability of the metric developed here could be investigated

Limitations and Future Analysis cont.

- ➔ If the dataset included win probability, and again was large enough to filter, it would be interesting to see how the density curves differ in expected passing/running situations, 20%<WP<80% situations, etc.

Citations

- Riske, 2020: https://www.pff.com/news/nfl-pff-data-study-sack-artist-pass-rushers
- Yurko, Ventura, and Horowitz, 2018: https://arxiv.org/pdf/1802.00998.pdf
- Gelman and Hill, 2007: Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge, United Kingdom: Cambridge University Press.
- Bates et al., 2015: “Fitting linear mixed-effects models ¨ using lme4,” Journal of Statistical Software, 67, 1–48.

Thank you guys for hosting this competition. It’s a fantastic cause and I really enjoyed working with the data.

- Alex